Power series approximations for two-class generalized processor sharing systems

Joris Walraevens* J.S.H. van Leeuwaarden• Onno J. Boxma°

Abstract: We develop power series approximations for a discrete-time queueing system with two parallel queues and one processor. If both queues are non-empty, a customer of queue 1 is served with probability β and a customer of queue 2 is served with probability $1-\beta$. If one of the queues is empty, a customer of the other queue is served with probability 1. We first describe the generating function $U(z_1, z_2)$ of the stationary queue lengths in terms of a functional equation, and show how to solve this using the theory of boundary value problems. Then, we propose to use the same functional equation to obtain a power series for $U(z_1, z_2)$ in β. The first coefficient of this power series corresponds to the priority case $\beta = 0$, which allows for an explicit solution. All higher coefficients are expressed in terms of the priority case. Accurate approximations for the mean stationary queue lengths are obtained from combining truncated power series and Padé approximation.

1 Introduction

Consider a discrete-time queueing model with two parallel queues that share a single processor. If both queues are non-empty at the beginning of a slot, a customer of queue 1 is served with probability (w.p.) β and a customer of queue 2 is served w.p. $1-\beta$. If one of the queues is empty, a customer of the other queue is served w.p. 1. This type of processor sharing occurs naturally in systems where different types of customers compete for resources. In telecommunication systems with integrated services, for instance, delay-sensitive streaming traffic shares resources with elastic traffic. So, if we consider the traffic arriving at queue 2 to be the delay-sensitive traffic in our model, a small β is necessary to limit the delay of this type of traffic. The exact choice of β should depend on the requirements (in terms of delay, loss, throughput, etc.) of both types of traffic.

The number of customers arriving at queue j ($j = 1, 2$) during slot k is denoted by $a_{j,k}$. We assume that $\{a_{j,k}, k > 0\}$ forms a sequence of independent and identically distributed (i.i.d.) random variables. We denote the bivariate probability generating function (pgf) of $a_{1,k}$ and $a_{2,k}$ by $A(z_1, z_2) := \mathrm{E}[z_1^{a_{1,k}} z_2^{a_{2,k}}]$. The mean number of arrivals in queue j is denoted by λ_j. The customers from queue j need service for a geometrically distributed number of slots with mean $1/\mu_j$. Since the service policy is work conserving, the stability condition is naturally given by $\rho = \lambda_1/\mu_1 + \lambda_2/\mu_2 < 1$.

The above queueing system gives rise to a random walk on the two-dimensional lattice in the quarter plane. The bivariate pgf of the stationary queue length distribution, denoted by $U(z_1, z_2)$, can be described in terms

*Department of Telecommunications and Information Processing, Ghent University - UGent, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium. Postdoctoral fellow with the Fund for Scientific Research, Flanders (FWO-Vlaanderen). E-mail: jw@telin.UGent.be. The research was done during a stay of the author at the EURANDOM Research institute, and was supported by a travel grant of the FWO-Vlaanderen.

•Eindhoven University of Technology and EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. Supported by an NWO VENI grant. E-mail: j.s.h.v.leeuwaarden@tue.nl

°Eindhoven University of Technology and EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. E-mail: o.j.boxma@tue.nl. Part of the research was done in the framework of the European Network of Excellence Euro-FGI, and of the Dutch BRICKS project.

of a functional equation of the type

$$K(z_1, z_2)U(z_1, z_2) = K_{00}(z_1, z_2)U(0, 0) + K_{10}(z_1, z_2)U(z_1, 0) + K_{01}(z_1, z_2)U(0, z_2), \tag{1}$$

where K, K_{00}, K_{01} and K_{10} depend on the input functions and input parameters; cf. (3)-(6). Using certain zero-tuples of the kernel $K(z_1, z_2)$, one can determine the functions $U(z_1, 0)$ and $U(0, z_2)$ as solutions to a Riemann boundary value problem. This approach, developed in [10, 12] and surveyed in [9], is for the model at hand outlined in Appendix A. The obtained formal solution, however, requires considerable numerical efforts, including the numerical determination of a conformal mapping.

Alternative approaches exist. The most general method for obtaining the stationary distribution of nearest-neighbor random walks in the quarter plane is developed by Fayolle, Iasnogorodski and Malyshev since the early seventies and summarized in the seminal book [13]. Their method also starts from (1), but then builds on the analytic continuation of the functions $U(z_1, 0)$ and $U(0, z_2)$. The rather ingenious continuation method, again combined with boundary value problems, then leads to solutions for $U(z_1, z_2)$ that are valid in the entire complex plane. Also, the conformal mapping that is required for our approach in Appendix A can in many cases be replaced by the use of Weierstrass elliptic functions, or by a Fredholm integral equation (as for Examples 1 and 2 in Subsection 5.1); for more details see [13]. We choose not to pursue the approach in [13], but rather to resort to a more numerically-oriented approach.

Other approaches for analyzing two-dimensional queueing models include the uniformization technique [17], the compensation method [3], and the power series approximation (PSA), see for instance [6, 7, 16]. For a comparison of the approaches see [1]. PSA is based on power series expansions of steady-state probabilities as functions of a certain parameter of the system, usually the load ρ, and was introduced in [16]. By using the balance equations of the queueing system, the coefficients of the terms in the power series can be calculated iteratively. A disadvantage of this approach is the deterioration of the accuracy when ρ increases. We propose a novel version of PSA that differs from the conventional approach in two ways. Firstly, we construct a power series expansion for the bivariate pgf $U(z_1, z_2)$ directly from the functional equation (1). Secondly, we construct a power series in β rather than in ρ. This makes sense, since we are primarily interested in the results of our model for small values of β (and for all possible values of the load). Note further that the queueing system is symmetric in β in the sense that $\beta = 0$ means priority for queue 2, and $\beta = 1$ means priority for queue 1. This symmetry is not present for the parameter ρ. Therefore, our PSA approach leads to the most accurate approximations not only near $\beta = 0$ but also near $\beta = 1$ (by constructing the power series in $1 - \beta$). This symmetry furthermore helps us in the construction of good approximations for all β.

Our PSA approach can be summarized as follows. For $U(z_1, z_2; \beta) := U(z_1, z_2)$ we construct the power series

$$U(z_1, z_2; \beta) = \sum_{m=0}^{\infty} V_m(z_1, z_2)\beta^m, \tag{2}$$

and we outline a procedure to determine the functions V_m iteratively. The first term V_0 of this power series corresponds to the priority case $\beta = 0$, which is well studied and allows for an explicit solution, cf. (14). The second term V_1 provides a first-order correction to the priority case for small β. All higher terms V_m can be expressed in V_0.

A final remark concerns the chosen modeling of the service times. Although deterministic service times of exactly one slot come natural for discrete-time queueing systems, we have opted to extend it to geometrically distributed service times. This does not complicate the analysis significantly, while it allows us to derive results for the well-known continuous-time generalized processor sharing queueing system (see [12]) directly from the discrete-time results. This is accomplished by letting the slot length go to zero and by scaling the arrival and service processes.

The paper is outlined as follows. In Section 2, we construct the functional equation for $U(z_1, z_2)$. An expression for $U(z_1, z_2)$ in terms of the solution of a boundary value problem is presented in Appendix A. In Section 3 we present, as our main contribution, the PSA approach for iteratively solving the functional equation.

Approximations obtained from the PSA for the mean queue length are discussed in Section 4, along with some numerical validations in Section 5. In Section 6, we show how our discrete-time framework leads to results for the continuous-time counterpart. Some conclusions are presented in Section 7.

2 The functional equation

The length of queue j at the beginning of slot k is denoted by $u_{j,k}$, $j = 1, 2$. We assume that the customer in service belongs to the queue it arrived in. We have the following system equations relating $(u_{1,k+1}, u_{2,k+1})$ to $(u_{1,k}, u_{2,k})$:

- If $u_{1,k} = 0, u_{2,k} = 0$: $u_{j,k+1} = a_{j,k}$, $j = 1, 2$.
- If $u_{1,k} = 0, u_{2,k} > 0$: $u_{1,k+1} = a_{1,k}$ and

$$u_{2,k+1} = \begin{cases} u_{2,k} - 1 + a_{2,k}, & \text{w.p. } \mu_2, \\ u_{2,k} + a_{2,k} & \text{w.p. } 1 - \mu_2. \end{cases}$$

- If $u_{1,k} > 0, u_{2,k} = 0$: $u_{2,k+1} = a_{2,k}$ and

$$u_{1,k+1} = \begin{cases} u_{1,k} - 1 + a_{1,k}, & \text{w.p. } \mu_1, \\ u_{1,k} + a_{1,k} & \text{w.p. } 1 - \mu_1. \end{cases}$$

- If $u_{1,k} > 0, u_{2,k} > 0$:

$$(u_{1,k+1}, u_{2,k+1}) = \begin{cases} (u_{1,k} - 1 + a_{1,k}, u_{2,k} + a_{2,k}), & \text{w.p. } \beta\mu_1, \\ (u_{1,k} + a_{1,k}, u_{2,k} - 1 + a_{2,k}), & \text{w.p. } (1-\beta)\mu_2, \\ (u_{1,k} + a_{1,k}, u_{2,k} + a_{2,k}), & \text{w.p. } \beta(1-\mu_1) + (1-\beta)(1-\mu_2). \end{cases}$$

We define $U_k(z_1, z_2)$ as $\mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\right]$, for all k, and we translate the system equations into pgfs:

$$\begin{aligned} U_{k+1}(z_1, z_2) =& A(z_1, z_2) \left[U_k(0,0) + \left(1 - \mu_2 + \frac{\mu_2}{z_2}\right)(U_k(0, z_2) - U_k(0,0)) + \left(1 - \mu_1 + \frac{\mu_1}{z_1}\right) \right. \\ & \times (U_k(z_1, 0) - U_k(0,0)) + \left(\beta\left(1 - \mu_1 + \frac{\mu_1}{z_1}\right) + (1-\beta)\left(1 - \mu_2 + \frac{\mu_2}{z_2}\right)\right) \\ & \left. \times (U_k(z_1, z_2) - U_k(0, z_2) - U_k(z_1, 0) + U_k(0,0)) \right]. \end{aligned}$$

In steady-state, $U_k(z_1, z_2)$ and $U_{k+1}(z_1, z_2)$ can be replaced by $U(z_1, z_2)$. By letting $k \to \infty$ in the above equation, we find the functional equation (1) with kernel

$$K(z_1, z_2) = z_1 z_2 - [(1 - \beta\mu_1 - (1-\beta)\mu_2) z_1 z_2 + (1-\beta)\mu_2 z_1 + \beta\mu_1 z_2] A(z_1, z_2) \tag{3}$$

and

$$K_{00}(z_1, z_2) = [\beta\mu_2 z_1(z_2 - 1) + (1-\beta)\mu_1(z_1 - 1)z_2] A(z_1, z_2) \tag{4}$$

$$K_{10}(z_1, z_2) = (1-\beta)[\mu_2 z_1(z_2 - 1) - \mu_1(z_1 - 1)z_2] A(z_1, z_2) \tag{5}$$

$$K_{01}(z_1, z_2) = \beta[\mu_1(z_1 - 1)z_2 - \mu_2 z_1(z_2 - 1)] A(z_1, z_2). \tag{6}$$

The functional equation (1) relates $U(z_1, z_2)$ to $U(z_1, 0)$, $U(0, z_2)$ and $U(0, 0)$ and can be solved using the theory of boundary value problems. This is done in Appendix A for a generalization of (1) that includes the starting position and transient behavior.

3 The power series approximation

We introduce the notation $U(z_1, z_2; \beta) := U(z_1, z_2)$ to express that this bivariate pgf is a function of β. First, we rearrange (1) as

$$\begin{aligned} G(z_1, z_2)U(z_1, z_2; \beta) - G_{10}(z_1, z_2)U(z_1, 0; \beta) - G_{00}(z_1, z_2)U(0, 0; \beta) \\ = \beta \cdot G_{10}(z_1, z_2)[U(z_1, z_2; \beta) - U(0, z_2; \beta) - U(z_1, 0; \beta) + U(0, 0; \beta)], \end{aligned} \tag{7}$$

where

$$\begin{aligned} G(z_1, z_2) &= z_2 - A(z_1, z_2)(\mu_2 + (1 - \mu_2)z_2), \\ G_{10}(z_1, z_2) &= A(z_1, z_2)(\mu_2(z_2 - 1) - \mu_1(1 - z_1^{-1})z_2), \\ G_{00}(z_1, z_2) &= A(z_1, z_2)\mu_1(1 - z_1^{-1})z_2. \end{aligned} \tag{8}$$

Note that one of the difficulties in solving the functional equation (7) is that it comprises *both* boundary functions $U(0, z_2; \beta)$ and $U(z_1, 0; \beta)$. Our approach is based on the observation that only one of the two boundary functions appears at the left-hand side of (7).

We assume, for the moment, that $U(z_1, z_2; \beta)$ is an analytic function of β in a neighborhood of 0. We will argue later on that this assumption is valid; see Subsection 3.1 below and Appendix B. We can then represent $U(z_1, z_2; \beta)$ by the power series expansion (2) for all z_1 and z_2 in the unit disk. Substitution of (2) into (7) yields

$$\begin{aligned} &G(z_1, z_2) \sum_{m=0}^{\infty} V_m(z_1, z_2)\beta^m - G_{10}(z_1, z_2) \sum_{m=0}^{\infty} V_m(z_1, 0)\beta^m - G_{00}(z_1, z_2) \sum_{m=0}^{\infty} V_m(0, 0)\beta^m \\ &= G_{10}(z_1, z_2) \sum_{m=0}^{\infty} [V_m(z_1, z_2) - V_m(0, z_2) - V_m(z_1, 0) + V_m(0, 0)]\beta^{m+1}. \end{aligned}$$

Equating coefficients of corresponding powers of β at both sides results in the following functional equation for V_m:

$$G(z_1, z_2)V_m(z_1, z_2) = G_{10}(z_1, z_2)(V_m(z_1, 0) + P_{m-1}(z_1, z_2)) + G_{00}(z_1, z_2)V_m(0, 0), \tag{9}$$

for all $m \geq 0$, with

$$P_m(z_1, z_2) := V_m(z_1, z_2) - V_m(0, z_2) - V_m(z_1, 0) + V_m(0, 0),$$

for $m \geq 0$ and $P_{-1}(z_1, z_2) := 0$.

We shall now outline how to determine expressions for $V_m(z_1, z_2)$. For a certain fixed m, we assume that $P_{m-1}(z_1, z_2)$ is known and we want to express V_m in terms of P_{m-1}. One can prove by a generalization of Rouché's theorem [2] that $G(z_1, z_2)$ (equation (8)) has one zero in the unit disk of z_2 for an arbitrary z_1 in the unit disk. Denote this zero by $Y(z_1)$. It is uniquely defined in the unit disk as $G(z_1, Y(z_1)) = 0$ and $|Y(z_1)| < 1$. The implicit function theorem then says that $Y(z_1)$ is an analytic function in the unit disk. In fact, $Y(z_1)$ is the pgf of a random variable (see [23] for a similar example). Since $U(z_1, z_2)$ is analytic for all z_1 and z_2 in the unit disk, the $V_m(z_1, z_2)$ are as well. Therefore, the right-hand side of (9) should equal zero for $z_2 = Y(z_1)$. This gives

$$V_m(z_1, 0) = -\frac{G_{00}(z_1, Y(z_1))}{G_{10}(z_1, Y(z_1))}V_m(0, 0) - P_{m-1}(z_1, Y(z_1)). \tag{10}$$

Upon substituting (10) into (9) we obtain

$$V_m(z_1, z_2) = \frac{1}{G(z_1, z_2)}\Big[\frac{G_{00}(z_1, z_2)G_{10}(z_1, Y(z_1)) - G_{10}(z_1, z_2)G_{00}(z_1, Y(z_1))}{G_{10}(z_1, Y(z_1))}$$

$$\times V_m(0,0) + G_{10}(z_1,z_2)Q_{m-1}(z_1,z_2)\Big] \tag{11}$$

with $Q_{-1}(z_1,z_2) := 0$ and for $m \geq 0$,

$$\begin{aligned} Q_m(z_1,z_2) :=& P_m(z_1,z_2) - P_m(z_1,Y(z_1)) \\ =& V_m(z_1,z_2) - V_m(z_1,Y(z_1)) - V_m(0,z_2) + V_m(0,Y(z_1)). \end{aligned} \tag{12}$$

The last step in finding an expression for V_m in terms of P_{m-1} (or Q_{m-1}) is the calculation of $V_m(0,0)$. This constant is found from the normalization condition. Since $U(1,1;\beta) = 1$ for all β, it follows that $V_0(1,1) = 1$ and $V_m(1,1) = 0$ for all $m > 0$. Setting $z_1 = z_2 = 1$ in (11) and using that $Q_m(1,1) = 0$ for all $m \geq 0$, we find that $V_0(0,0) = 1-\rho$ and $V_m(0,0) = 0$ for $m > 0$. We finally arrive at the following relation between V_m and Q_{m-1} for $m > 0$:

$$V_m(z_1,z_2) = \frac{A(z_1,z_2)[\mu_2 z_1(z_2-1) - \mu_1(z_1-1)z_2]Q_{m-1}(z_1,z_2)}{z_1[z_2 - A(z_1,z_2)(\mu_2 + (1-\mu_2)z_2)]}. \tag{13}$$

Here,

$$V_0(z_1,z_2) = \frac{\mu_1\mu_2(1-\rho)A(z_1,z_2)(z_1-1)(z_2-Y(z_1))}{[z_2 - A(z_1,z_2)(\mu_2+(1-\mu_2)z_2)][\mu_1(z_1-1)Y(z_1) - \mu_2 z_1(Y(z_1)-1)]}. \tag{14}$$

Hence, starting from V_0 in (14), every function V_m can be determined iteratively via (13) and (12).

For $\beta = 0$, the second queue has strict priority over the first. The pgf of the queue length in this case equals $U(z_1,z_2;0) = V_0(z_1,z_2)$ as given in (14). This last expression is indeed the pgf in a discrete-time preemptive resume priority queueing system with geometric service times and the first queue having low priority (see e.g. [24]).

3.1 Analyticity of $U(z_1,z_2,\beta)$ in a neighborhood of $\beta = 0$

Proving the analyticity of $U(z_1,z_2,\beta)$ in a neighborhood of $\beta = 0$ is not straightforward. There are however several (potential) approaches to tackle this problem.

A first approach is to combine the implicit function theorem with the balance equations of the Markov chain that describes the queue lengths. The implicit function theorem basically says that if a set of N equations in $N+1$ variables $(x_1,,\ldots,x_N,y)$ satisfies certain conditions, and if $(x_1^{(0)},\ldots,x_N^{(0)},y^{(0)})$ is a known solution of the set of equations, then there are unique (analytic) functions $u_1(y),\ldots,u_N(y)$ such that $u_i(y^{(0)}) = x_i^{(0)}$ $(i = 1,\ldots,N)$ and $(u_1(y),\ldots,u_N(y),y)$ is a solution of the set of equations (see for instance [19]). In our problem, x_i would represent the mass functions of the queue lengths and y would be β. This approach is for instance taken in [16] for a related problem. The main difficulty is that the implicit function theorem only applies to a *finite* set of equations. Therefore, in [16], the proof consists of three steps: (i) construction of a finite Markov chain from the original infinite Markov chain, (ii) use of the implicit function theorem to prove that the stationary process of this Markov chain is analytic in the parameter and (iii) a proof that this analyticity is carried over to the (stationary process of the) infinite Markov chain. A fourth step in our case should be a proof that the analyticity of the probability mass function of the queue lengths of both queues is carried over to the bivariate pgf $U(z_1,z_2)$.

A second approach would be to reformulate (1) in terms of a Dirichlet problem for a circle, for which Schwarz's formula (see [8]) yields the solution to the functional equation in terms of an elliptic integral. Such an integral would provide an explicit characterization, and for specific choices of $A(z_1,z_2)$ one could then investigate the analyticity of $U(z_1,z_2,\beta)$. See also Remark 7 at the end of appendix B.

A third approach is the use of the implicit function theorem on the functional equation (1) directly. For this problem the 'traditional' implicit function theorem is of no use, since we do not have a set of equations. However, there are variants for the implicit function theorem in functional analysis (see for instance [11]).

For our problem, the proof could go as follows: the left hand side of the functional equation (1) is a mapping $f(\beta, U(z_1, z_2))$ from the product space of two Banach spaces to another Banach space. If this mapping satisfies certain conditions (most notably that the functional derivative $D_2(f)(\beta, U(z_1, z_2))$ is a linear homeomorphism) and if there is a $U_0(z_1, z_2)$ so that $f(0, U_0(z_1, z_2)) = 0$ then there exists a unique (analytic) function $W(z_1, z_2, \beta)$ so that $f(\beta, W(z_1, z_2, \beta)) = 0$. In our case, we have a solution $U_0(z_1, z_2) = V_0(z_1, z_2)$, see expression (14).

Yet another approach is the one discussed in the PhD thesis of W.B. van den Hout ([22], Section 2.5). Van den Hout first shows that the power series in, say, β as obtained in the PSA do not necessarily converge for all β inside the unit circle, and may not even be analytic at $\beta = 0$. He then provides assumptions which are sufficient for such analyticity in a neighbourhood of $\beta = 0$; see in particular Theorem 2.1 on p. 41 of [22]. Translated into our setting, we would have to pose conditions on $A(z_1, z_2)$ such that (i) one can bound $V_m(z_1, z_2)$ from above by some $\hat{V}_m(z_1, z_2)$ for all $|z_1| \leq 1$, $|z_2| \leq 1$ and for $m > M$ for some $M > 0$, and (ii) $\sum_m \beta^m \hat{V}_m(z_1, z_2)$ converges in a neighbourhood of $\beta = 0$. The latter holds in particular if the $\hat{V}_m(z_1, z_2)$ decay geometrically fast in m. The recursive expression (13) of $V_m(z_1, z_2)$ in $Q_{m-1}(z_1, z_2)$, and hence in $V_{m-1}(z_1, z_2)$, is the key to proving that bound for $V_m(z_1, z_2)$.

We have chosen the third approach to prove the analyticity of U in a neighborhood of $\beta = 0$. A sketch of this proof is given in Appendix B.

3.2 Conservation laws

Let us first look at the special case $\mu = \mu_1 = \mu_2$. The pgf of the total number of customers in the queue can be found from (1) by setting $z = z_1 = z_2$:

$$U(z, z) = \frac{A(z, z)\mu(z - 1)(1 - \rho)}{z - A(z, z)(\mu + (1 - \mu)z)}. \tag{15}$$

Expression (15) for $U(z, z)$ is in fact the pgf of a discrete-time $M^X/Geo/1$ queue with the pgf of the number of arrivals in a slot equal to $A(z, z)$ and with geometrical service times with mean $1/\mu$. It is clear that the total system behaves as such a queueing system when $\mu_1 = \mu_2$, since customers are served with rate μ when the system is busy, irrespective of β. As a result, $V_0(z, z) = U(z, z)$ and $V_m(z, z) = 0$ for $m \geq 1$, which also follows from the PSA approach.

When $\mu_1 \neq \mu_2$, the total system content is no longer independent of β. However, the total amount of unfinished work expressed in number of slots does not depend on β. Since each customer of queue j present in the system at the beginning of a slot needs a geometrically distributed number of slots service time with mean $1/\mu_j$, the pgf of the total unfinished work is given by:

$$U\left(\frac{\mu_1 z}{1 - (1 - \mu_1)z}, \frac{\mu_2 z}{1 - (1 - \mu_2)z}\right) = \frac{A\left(\frac{\mu_1 z}{1 - (1 - \mu_1)z}, \frac{\mu_2 z}{1 - (1 - \mu_2)z}\right)(z - 1)U(0, 0)}{z - A\left(\frac{\mu_1 z}{1 - (1 - \mu_1)z}, \frac{\mu_2 z}{1 - (1 - \mu_2)z}\right)},$$

which is again found from (1). So we have that

$$V_0\left(\frac{\mu_1 z}{1 - (1 - \mu_1)z}, \frac{\mu_2 z}{1 - (1 - \mu_2)z}\right) = U\left(\frac{\mu_1 z}{1 - (1 - \mu_1)z}, \frac{\mu_2 z}{1 - (1 - \mu_2)z}\right)$$

and

$$V_m\left(\frac{\mu_1 z}{1 - (1 - \mu_1)z}, \frac{\mu_2 z}{1 - (1 - \mu_2)z}\right) = 0, \quad m > 0.$$

This is also found from expressions (13) and (14). We will make use of this property when approximating the mean number of customers in the queues.

4 Approximations of performance measures

We have obtained an algorithm to determine the $V_m(z_1, z_2)$ for each desired m, but the actual calculation is far from straightforward. The reason for this is that l'Hôpital's rule has to be used multiple times, which leads to expressions for $V_m(z_1, z_2)$ that become more complex with m. For instance, determining $Q_m(z_1, z_2)$ requires $V_{m-1}(z_1, Y(z_1))$ and $V_{m-1}(0, z_2)$. The application of l'Hôpital's rule is necessary to obtain these latter functions from (13). This problem is even more significant when calculating performance measures such as the moments; the generating functions then have to be differentiated in 1 which again requires multiple applications of l'Hôpital's rule.

We restrict the remaining discussion to the mean queue length in queue j given by

$$\mathrm{E}[u_j] = \sum_{m=0}^{\infty} \beta^m \frac{\partial V_m(z_1, z_2)}{\partial z_j}\Big|_{z_1=z_2=1}. \tag{16}$$

Since $V_m(\mu_1 z/(1-(1-\mu_1)z), \mu_2 z/(1-(1-\mu_2)z)) = 0$ for $m > 0$, it is easily seen that

$$\frac{\partial V_m(z_1, z_2)}{\partial z_1}\Big|_{z_1=z_2=1} = -\frac{\mu_1}{\mu_2}\frac{\partial V_m(z_1, z_2)}{\partial z_2}\Big|_{z_1=z_2=1},$$

for $m \geq 1$. Therefore, for $m \geq 1$, we only need to calculate one of the two derivatives in the previous formula. The partial derivative $\partial V_0(z_1, z_2)/\partial z_j|_{z_1=z_2=1}$, $j = 1, 2$ can be calculated easily from (14). We get the results for the mean queue lengths in the low-priority and high-priority queue in the discrete-time $M^X/Geo/1$ preemptive resume priority queue, as discussed before.

Let us now assume to have found the exact values $v_{j,m}$ for $\partial V_m(z_1, z_2)/\partial z_j|_{z_1=z_2=1}$ for $j = 1, 2$ and $m = 0, \ldots, M$. Truncation of the power series (16) leads to

$$\mathrm{E}[u_j] = \sum_{m=0}^{M} v_{j,m}\beta^m + O(\beta^{M+1}). \tag{17}$$

This truncation yields accurate approximations for small β.

The problem is symmetric in β in the sense that the PSA can also be constructed in $\bar{\beta} = 1 - \beta$ (instead of in β). If the $v_{j,m}$ are calculated for general $A(z_1, z_2)$, μ_1 and μ_2, this second approximation can also be calculated directly by interchanging the roles of both queues. So, if one is interested in $\mathrm{E}[u_j]$ for β near 1, we can use that

$$\mathrm{E}[u_j] = \sum_{m=0}^{M} \tilde{v}_{j,m}(1-\beta)^m + O((1-\beta)^{M+1}), \tag{18}$$

with $\tilde{v}_{j,m} = \partial \tilde{V}_m(z_1, z_2)/\partial z_{3-j}|_{z_1=z_2=1}$, $\tilde{V}_m$ given by V_m (in (13)) with $A(z_1, z_2)$ replaced by $A(z_2, z_1)$, and μ_1 and μ_2 interchanged.

Truncation yields approximations which are accurate near 0 or 1. In fact they provide the exact 0- to M-th order derivatives in 0 or 1. Padé approximants replace the power series (16) by a rational functional. We can hence approximate $\mathrm{E}[u_j]$ by

$$[L/N]_{E[u_j]}(\beta) = \frac{\sum_{l=0}^{L} u_{j,l}\beta^l}{\sum_{n=0}^{N} w_{j,n}\beta^n}, \tag{19}$$

with L, N and the coefficients $u_{j,l}$ and $w_{j,n}$ chosen such that

$$[L/N]_{E[u_j]}(\beta) = \sum_{m=0}^{M} v_{j,m}\beta^m + O(\beta^{M+1}),$$

$$[L/N]_{E[u_j]}(\beta) = \sum_{m=0}^{M} \tilde{v}_{j,m}(1-\beta)^m + O((1-\beta)^{M+1}),$$

i.e., such that the approximant has the correct derivatives up to order M in both 0 and 1. These $2(M+1)$ derivatives to be matched by expression (19) lead to a set of $2(M+1)$ equations with $L+N+1$ unknowns (coefficients $u_{j,l}$ ($l=0,\ldots,L$) and $w_{j,n}$ ($n=1,\ldots,N$); we use the normalization $w_{j,0}=1$). Every choice of (L,N) with $L+N=2M+1$ thus leads in general to a unique solution for the $u_{j,l}$ ($l=0,\ldots,L$) and $w_{j,n}$ ($n=1,\ldots,N$) in terms of $v_{j,m}$ and $\tilde{v}_{j,m}$, $m=0,\ldots,M$. If one is interested in approximate formulas which are accurate for the whole range $[0,1]$ of β, the generalized Padé approximants are the best choice. Note that the $[2M+1/0]_{E[u_j]}$ approximant is a polynomial, like the truncated version of the PSAs. We have observed that the $[0/2M+1]_{E[u_j]}$ approximant is usually among the most accurate ones (see also Section 5). Note though that these Padé approximants should be used carefully, since the denominators can introduce poles in the approximation.

Remark 1. (First-order correction) The second term V_1 in the PSA equals the difference between the model described in this paper and the preemptive resume priority queue, for β going to 0. We find the following limits:

$$\begin{aligned}
\lim_{\beta\to 0}\frac{U(z_1,z_2;\beta)-U(z_1,z_2;0)}{\beta} =& \frac{A(z_1,z_2)[\mu_2 z_1(z_2-1)-\mu_1(z_1-1)z_2]}{z_1[z_2-A(z_1,z_2)(\mu_2+(1-\mu_2)z_2)]}\\
&\times[U(z_1,z_2;0)-U(z_1,Y(z_1);0)-U(0,z_2;0)+U(0,Y(z_1);0)]\\
\lim_{\beta\to 0}\frac{\mathrm{E}[u_2]-\mathrm{E}[u_2|\beta=0]}{\beta} =& \frac{\mathrm{E}[u_2|\beta=0]-\mathrm{E}[u_2 1_{u_1=0}|\beta=0]}{1-\rho_2}\\
\lim_{\beta\to 0}\frac{\mathrm{E}[u_1]-\mathrm{E}[u_1|\beta=0]}{\beta} =& -\frac{\mu_1}{\mu_2}\frac{\mathrm{E}[u_2|\beta=0]-\mathrm{E}[u_2 1_{u_1=0}|\beta=0]}{1-\rho_2},
\end{aligned} \tag{20}$$

with 1_X the indicator function of the event X. These limits are important as they give first-order correction terms to the priority results for a near-priority queueing system.

5 Numerical examples

We now compare the PSA approximations to simulation results and investigate the influence of some parameters on the mean queue lengths. Throughout this section, we consider deterministic service times of one slot. We consider a generalized processor sharing discipline as analysed in the paper. Thus, when both queues are non-empty at the beginning of a slot, a customer of queue 1 (queue 2) is served w.p. β ($1-\beta$) during that slot. If one of the queues is empty, the other queue is served. Because of the work-conservation property, we can concentrate on the mean queue length of only one queue, say queue 2.

5.1 Validation of the approximations

Example 1. Assume the number of arrivals to both queues and the total number of arrivals in a slot to be binomially distributed. More precisely, assume the following bivariate pgf of the number of class-1 and class-2 arrivals during a slot:

$$A(z_1,z_2)=\left(1+\frac{\lambda_1}{2}(z_1-1)+\frac{\lambda_2}{2}(z_2-1)\right)^2. \tag{21}$$

Figure 1 depicts the approximations (17) and (18) as a function of β for increasing M. The arrival rates λ_1 and λ_2 are 0.7 and 0.1. We have simulated the system for $\beta=0,0.1,\ldots,1$ (crosses in the figure). The horizontal lines ($M=0$) equal the values for the priority queues ($\beta=0$ and $\beta=1$ respectively). Figure 1 confirms that the PSA approximations are indeed accurate for β near 0 and near 1 and that more terms provide larger regions for β where the accuracy is good. However, as can also be seen from this figure, the approximations deteriorate (rapidly) for β away from 0 and 1. This can be dealt with using the Padé approximants introduced in Section 4. Figure 2 depicts the Padé approximants (19) for the same example as in Figure 1 and for $M=3$. It displays the approximants for $N=0,2,5$ and 7 (and $L=2M+1-N$).

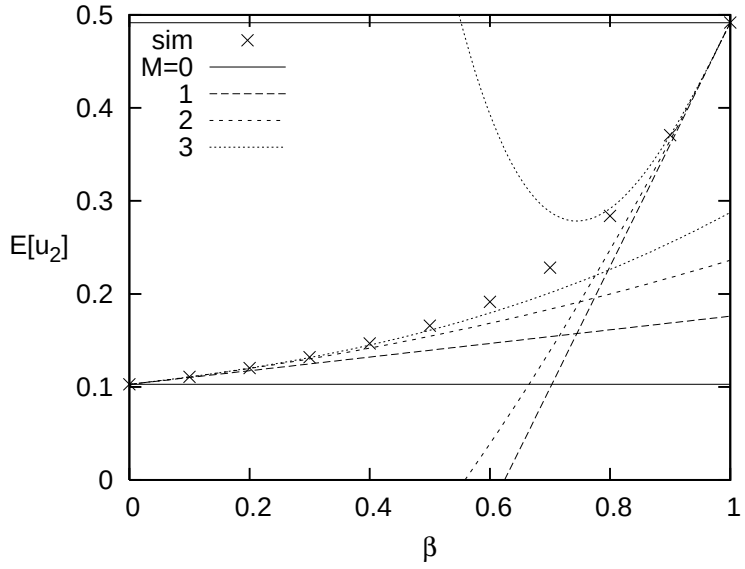

Figure 1: Truncation approximations of $\mathrm{E}[u_2]$ for binomially distributed arrival batch sizes with arrival rates $\lambda_1 = 0.7$ and $\lambda_2 = 0.1$.

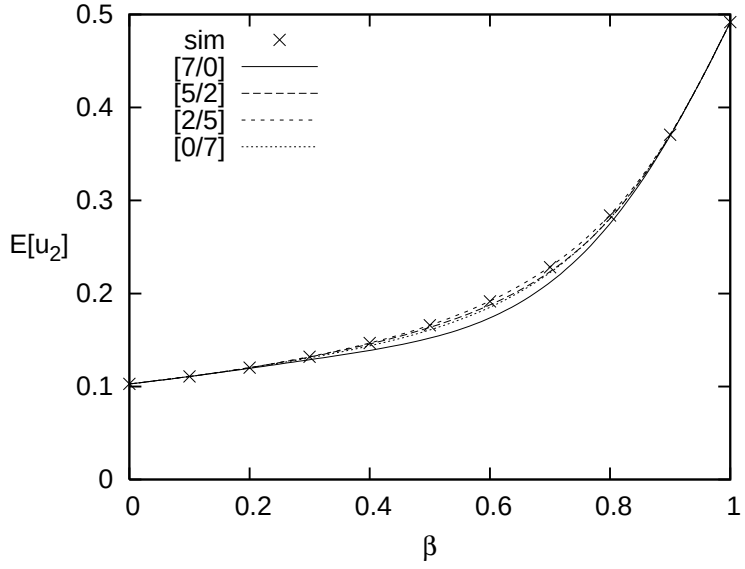

Figure 2: Padé approximations of $\mathrm{E}[u_2]$ for binomially distributed arrival batch sizes with arrival rates $\lambda_1 = 0.7$ and $\lambda_2 = 0.1$.

Example 2. Assume the arrivals of both classes to be a sequence of two independent geometrically distributed random variables with means λ_1 and λ_2 respectively, i.e.

$$A(z_1, z_2) = \frac{1-\lambda_1}{1-\lambda_1 z_1} \frac{1-\lambda_2}{1-\lambda_2 z_2}.$$

We again concentrate on queue 2. Figure 3 depicts the mean class-2 content as a function of β for $\lambda_1 = 0.7$ and $\lambda_2 = 0.1$. The same conclusions as for Figure 1 can be drawn, albeit that it is more pronounced that the PSA for β works best for small β and the one for $\bar{\beta} = 1 - \beta$ for β near 1. We can again construct similar Padé approximants.

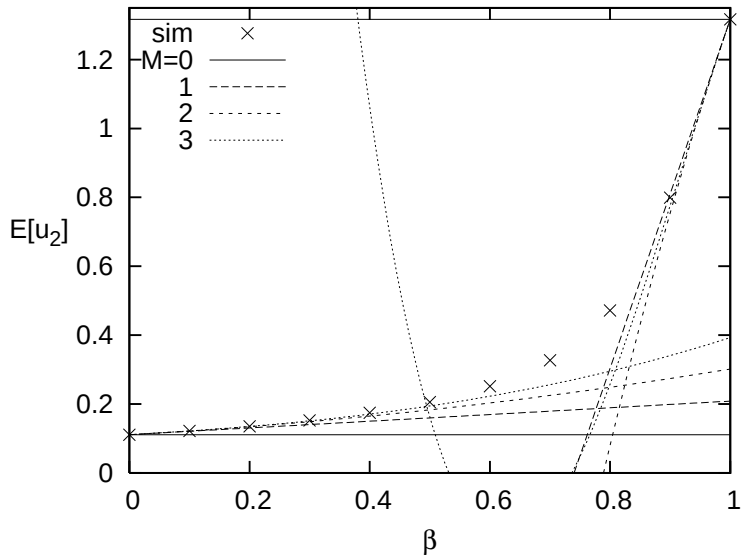

Figure 3: Truncation approximations of $\mathrm{E}[u_2]$ for geometrically distributed arrival batch sizes with means $\lambda_1 = 0.7$ and $\lambda_2 = 0.1$.

λ \ α	0.1	0.3	0.5	0.7	0.9
0.1	4.8334e-4	1.1755e-3	1.4608e-3	1.2829e-3	5.7582e-4
0.2	2.0833e-3	5.2967e-3	6.9146e-3	6.4153e-3	3.0622e-3
0.3	5.0689e-3	1.3516e-2	1.8668e-2	1.8533e-2	9.6076e-3
0.4	9.7846e-3	2.7456e-2	4.0439e-2	4.3625e-2	2.5293e-2
0.5	1.6678e-2	4.9428e-2	7.8298e-2	9.3511e-2	6.3182e-2
0.6	2.6339e-2	8.2767e-2	1.4231e-1	1.9231e-1	1.6054e-1
0.7	3.9567e-2	1.3237e-1	2.4944e-1	3.9111e-1	4.3844e-1
0.8	5.7460e-2	2.0558e-1	4.2880e-1	8.0254e-1	1.3614e+0
0.9	8.1576e-2	3.1367e-1	7.3150e-1	1.6875e+0	5.1812e+0

Table 1: First-order corrections for near-priority queues

5.2 Influence of input parameters on the mean system content for near-priority systems

Again consider the example of the binomially distributed arrival batch sizes (Expression (21)) and service times of one slot. The total arrival rate is given by $\lambda = \lambda_1 + \lambda_2$ and we define the fraction of class-2 arrivals $\alpha := \lambda_2/\lambda$. Table 1 displays some values of the first-order correction term (the first derivative of the mean class-2 queue content in $\beta = 0$, see Formula (20)) for particular values of λ and α. Some interesting observations can be made: Firstly, the correction term increases with increasing load for a constant value of α. This is expected since for high loads the number of customers in the class-1 buffer is usually non-zero, thus sharing the server with this buffer, even for a small percentage of the time can have a large influence. For small arrival rates, one of the buffers is almost always empty and the service discipline does not play an important role. A second observation is that for a given arrival rate the correction term first increases with α, reaches a maximum and then decreases. In $\alpha = 0$ and $\alpha = 1$, the correction term equals 0.

Similar conclusions can be drawn from Table 2, where we show the relative first-order correction

$$\frac{\partial V_1(z_1, z_2)}{\partial z_1} \Bigg/ \left.\frac{\partial V_0(z_1, z_2)}{\partial z_1}\right|_{z_1=z_2=1}.$$

λ \ α	0.1	0.3	0.5	0.7	0.9
0.1	4.8213e-2	3.8883e-2	2.8836e-2	1.7989e-2	6.2436e-3
0.2	1.0363e-1	8.6892e-2	6.7277e-2	4.4032e-2	1.6127e-2
0.3	1.6767e-1	1.4656e-1	1.1920e-1	8.2751e-2	3.2572e-2
0.4	2.4209e-1	2.2126e-1	1.9030e-1	1.4200e-1	6.1596e-2
0.5	3.2922e-1	3.1560e-1	2.8910e-1	2.3548e-1	1.1656e-1
0.6	4.3210e-1	4.3589e-1	4.2847e-1	3.8770e-1	2.2984e-1
0.7	5.5480e-1	5.9106e-1	6.2813e-1	6.4360e-1	4.8814e-1
0.8	7.0296e-1	7.9392e-1	9.1885e-1	1.0872e+0	1.1509e+0
0.9	8.8453e-1	1.0634e+0	1.3495e+0	1.8789e+0	3.0964e+0

Table 2: Relative first-order corrections for near-priority queues

This is a measure for the relative effect of giving a small share of the processor's time to the class-1 queue in the priority system. Here, the relative first-order correction does not go to 0 for $\alpha \to 0$ (in fact it tends to $\lambda/(2-\lambda)$ for this arrival process). The reason is that the mean class-2 queue length in the priority case equals 0 for $\alpha = 0$. For low loads, the value of the relative first-order correction term is maximal for $\alpha = 0$ and is a strictly decreasing function in α. For high loads, a maximum is reached for some $\alpha > 0$.

We can conclude that an increase of the total arrival rate not only results in an absolute increase of the first-order correction term, but also in an increase relative to the priority result. A GPS schedule has more impact when the arrival rate is high. Our results are therefore especially useful for these high arrival rates as they give a significant first-order (and even higher-order) correction term(s) to the priority result. The examples also show that these correction terms are sensitive to the parameters of the system, which in turn shows the necessity of the obtained formulas.

6 Continuous-time results

We now sketch how the discrete-time results lead to results for the continuous-time generalized processor sharing system. This continuous-time model is the most prominent subclass of the model in [12], there called coupled processors; see also [18] for an analysis of the symmetric coupled processor model and see [20] for an approximate analysis of "cycle stealing" in coupled processors.

Assume that arrivals occur to both queues according to independent Poisson processes with arrival rates λ_1^* and λ_2^* respectively. The service times of customers in queue j are exponentially distributed with mean $1/\mu_j^*$. The processor is shared in the following way: when both queues are non-empty, queue 1 is served with rate β ($0 \leq \beta \leq 1$ without loss of generality) while queue 2 is served with rate $1-\beta$. When one of both queues is empty, the other queue is served with rate 1.

We outline in this section that we can find the distribution of the number of customers at a random point in time from the results of the discrete-time case. We therefore show that the pgfs of the numbers of arrivals and the service times as well as the functional equation (7) go to their continuous-time counterparts. Similar approaches were taken in [4, 21].

Let us divide the time axis into equal intervals of length Δ. We first define arrival and service processes in the discrete-time case that scale to the above continuous-time arrival and service processes. Define therefore

$$\mu_j = \mu_j^* \Delta, \tag{22}$$

$$A(z_1, z_2) = (1 - \lambda_1^* \Delta + \lambda_1^* \Delta z_1)(1 - \lambda_2^* \Delta + \lambda_2^* \Delta z_2). \tag{23}$$

The service times of class-j expressed in number of slots are geometrically distributed with parameter μ_j, so

their pgf is given by

$$S_j(z) = \frac{\mu_j^* \Delta}{1 - (1 - \mu_j^* \Delta) z}. \tag{24}$$

The Laplace-Stieltjes transform of the service times in continuous time for the slot lengths going to zero equals

$$S_j^*(s) = \lim_{\Delta \to 0} S_j(e^{-s\Delta}) = \frac{\mu_j^*}{s + \mu_j^*}. \tag{25}$$

The interarrival times between non-empty batches in the discrete-time model are geometrically distributed with parameter $A(0,0)$, i.e., their pgf is given by

$$I(z) = \frac{(1 - A(0,0))z}{1 - A(0,0)z},$$

with $A(0,0) = 1 - (\lambda_1^* + \lambda_2^*)\Delta + O(\Delta^2)$. Similarly as for the service times, we get that the Laplace-Stieltjes transform of the interarrival times in continuous time for the slot length going to zero is exponential with mean $1/(\lambda_1^* + \lambda_2^*)$. The pgf of the number of customers arriving in a batch is then given by

$$\frac{A(z_1, z_2) - A(0,0)}{1 - A(0,0)} = \frac{(\lambda_1^* z_1 + \lambda_2^* z_2)\Delta + O(\Delta^2)}{(\lambda_1^* + \lambda_2^*)\Delta + O(\Delta^2)}.$$

Hence, for Δ going to zero, the interarrival times between batches are exponentially distributed with mean $1/(\lambda_1^* + \lambda_2^*)$ and a batch consists of a customer arrival in the first queue with probability $\lambda_1^*/(\lambda_1^* + \lambda_2^*)$ or a customer arrival in the second queue with probability $\lambda_2^*/(\lambda_1^* + \lambda_2^*)$. We thus conclude that the arrival process into both queues converges to two independent Poisson processes with parameters λ_1^* and λ_2^*.

With μ_j and $A(z_1, z_2)$ as in (22)-(23), the discrete-time arrival and service processes converge to Poisson arrivals and exponential service times for $\Delta \to 0$. Since β is the probability that a class-1 customer is served in a slot, this is the fraction of time class-1 customers are served when both queues are non-empty, in the limit for Δ going to zero. When we substitute μ_j and $A(z_1, z_2)$ by (22)-(23) in the functional equation (7), we arrive at

$$F_1(z_1, z_2)\Delta + F_2(z_1, z_2)\Delta^2 + F_3(z_1, z_2)\Delta^3 = 0, \tag{26}$$

with

$$F_1(z_1, z_2) = K^*(z_1, z_2)U(z_1, z_2) - K_{00}^*(z_1, z_2)U(0,0) - K_{10}^*(z_1, z_2)U(z_1, 0) - K_{01}^*(z_1, z_2)U(0, z_2)$$

and

$$\begin{aligned}
K^*(z_1, z_2) =& (1 - \beta)\mu_1(z_1 - 1)z_2 + \beta\mu_2 z_1(z_2 - 1) - z_1 z_2(\lambda_1(z_1 - 1) - \lambda_2(z_2 - 1)) \\
K_{00}^*(z_1, z_2) =& \beta\mu_1(z_1 - 1)z_2 + (1 - \beta)\mu_2 z_1(z_2 - 1) \\
K_{10}^*(z_1, z_2) =& \beta(\mu_2 z_1(z_2 - 1) - \mu_1(z_1 - 1)z_2) \\
K_{01}^*(z_1, z_2) =& (1 - \beta)(\mu_1(z_1 - 1)z_2 - \mu_2 z_1(z_2 - 1)).
\end{aligned}$$

The precise expressions of $F_2(z_1, z_2)$ and $F_3(z_1, z_2)$ are not important for the further discussion. It suffices to know that - like $F_1(z_1, z_2)$ - they are linear functions in $U(z_1, z_2)$, $U(z_1, 0)$, $U(0, z_2)$ and $U(0, 0)$, with coefficients analytic in the whole complex plane. Therefore, the three functions F_i are analytic at least in the region $|z_1| < 1$ and $|z_2| < 1$. For $\Delta \to 0$ the first term in (26) is dominant. The functional equation thus converges to

$$F_1(z_1, z_2) = 0 \tag{27}$$

for $\Delta \to 0$. This is indeed the functional equation of the continuous-time GPS system described in [12]. Since there is only one normalized solution of this functional equation inside the unit disk, the solution described in this paper for the discrete-time functional equation evolves to the solution of (27) for $\Delta \to 0$ and we can thus directly find approximations for the continuous-time GPS queueing system from the results in sections 3-4.

Note that we can also find results for the continuous-time GPS system with batch arrivals, by choosing a less restrictive $A(z_1, z_2)$. Our discrete-time results can furthermore show the influence of discretizing time on the performance measures. When a continuous-time model is used to approximate a discrete-time model, our results can also quantify the error that is introduced, as a function of the slot length.

7 Conclusions

We studied a discrete-time two-queue Generalized Processor Sharing system in this paper. When customers are present in both queues, the queues are served with probability β and $1-\beta$ respectively. We developed a novel technique based on Power Series Approximations of the joint probability generating function in the parameter β. The coefficients of the power terms are iteratively calculated starting from the constant term. This constant term is the joint pgf of a priority queue ($\beta = 0$).

By truncating the power series, we find good approximations for the means of the numbers of customers in both queues, for small β and, by symmetry, for β near 1. Interpolation techniques lead to more accurate approximations. A major advantage of the technique over the standard boundary value problem solution technique (described in the appendix) is that the formulas for say the mean queue lengths are explicit in the input parameters and require no additional numerical effort. Therefore they can be important in for instance control problems, where an optimal β is to be found given the delay requirements of both types of traffic.

The developed technique is promising to deal in general with the analysis of queueing systems with some sort of coupling. Examples are queues with a Packet-based Generalized Processor Sharing scheduling and a tandem queue where the two queues share a single processor. Another interesting topic is the Generalized Processor Sharing queue with *three* (or more) classes. The theory of boundary value problems has not been developed for problems with more than two dimensions.

A Solution to the boundary value problem

In this appendix, we present the solution to a more general form of the functional equation (1) in terms of a Riemann-Hilbert boundary value problem. Introducing $U_{k;x,y}(z_1, z_2) := U_k(z_1, z_2)$ with initial condition $u_{1,0} = x,\ u_{2,0} = y$ and

$$\Phi_{x,y}(r, z_1, z_2) := \sum_{k=0}^{\infty} r^k U_{k;x,y}(z_1, z_2), \tag{28}$$

we have for $|r| < 1, |z_1| \leq 1, |z_2| \leq 1$:

$$\begin{aligned}[z_1 z_2 - r\psi(z_1, z_2)]\Phi_{x,y}(r, z_1, z_2) =& rA(z_1, z_2)\{[\beta\mu_2 z_1(z_2 - 1) + (1-\beta)\mu_1(z_1 - 1)z_2]\Phi_{x,y}(r, 0, 0) \\ &+ (1-\beta)[\mu_2 z_1(z_2 - 1) - \mu_1(z_1 - 1)z_2]\Phi_{x,y}(r, z_1, 0) \\ &+ \beta[\mu_1(z_1 - 1)z_2 - \mu_2 z_1(z_2 - 1)]\Phi_{x,y}(r, 0, z_2)\} + z_1^{x+1} z_2^{y+1}, \end{aligned} \tag{29}$$

with

$$\psi(z_1, z_2) := A(z_1, z_2)[(1 - \beta\mu_1 - (1-\beta)\mu_2)z_1 z_2 + (1-\beta)\mu_2 z_1 + \beta\mu_1 z_2]. \tag{30}$$

It should be noticed that $\psi(z_1, z_2)$ is the generating function of a pair of non-negative, integer-valued random variables. Formula (29) has the following global form:

$$\begin{aligned} &K(r, z_1, z_2)\Phi_{x,y}(r, z_1, z_2) \\ &\quad = rK_{00}(z_1, z_2)\Phi_{x,y}(r, 0, 0) + rK_{10}(z_1, z_2)\Phi_{x,y}(r, z_1, 0) + rK_{01}(z_1, z_2)\Phi_{x,y}(r, 0, z_2) + z_1^{x+1} z_2^{y+1}, \end{aligned} \tag{31}$$

with the *kernel* $K(r, z_1, z_2)$ being defined as

$$K(r, z_1, z_2) := z_1 z_2 - r\psi(z_1, z_2). \tag{32}$$

Random walks on the two-dimensional lattice in the first quadrant of the plane that give rise to functional equations and kernels of the type as in (31) and (32) are discussed by Cohen [9]. In such random walks, the steps to the West, South-West and South are at most one. In [9], it is sketched how this class of random walks, typically arising in queueing models, can be analysed via a transformation to a two-dimensional boundary value problem of mathematical physics, like a Riemann or Riemann-Hilbert problem. In Part II of [10], a much more detailed exposition of this approach is presented, for a slightly more restricted class of two-dimensional random walks. The kernel of the random walk that is studied there contains the kernel that features in (32). In this appendix, we sketch the way in which this boundary value approach can be used to determine the generating function $\Phi_{x,y}(r, z_1, z_2)$ of the two-dimensional queueing problem that was presented in Section 1. We distinguish between four steps.

A.1 The zerotuples of the kernel

Obviously, the generating function $\Phi_{x,y}(r, z_1, z_2)$ should be finite for all zerotuples $(\hat{z}_1, \hat{z}_2)$ of $K(r, z_1, z_2)$ with $|\hat{z}_1| \leq 1$, $|\hat{z}_2| \leq 1$.
Let

$$A(r) := \{(z_1, z_2) : K(r, z_1, z_2) = 0,\ |z_1| \leq 1,\ |z_2| \leq 1\}.$$

Then for all $(\hat{z}_1, \hat{z}_2) \in A$, one must require that

$$rK_{00}(\hat{z}_1, \hat{z}_2)\Phi_{x,y}(r, 0, 0) + rK_{10}(\hat{z}_1, \hat{z}_2)\Phi_{x,y}(r, \hat{z}_1, 0) + rK_{01}(\hat{z}_1, \hat{z}_2)\Phi_{x,y}(r, 0, \hat{z}_2) + \hat{z}_1^{x+1}\hat{z}_2^{y+1} = 0. \tag{33}$$

If one can construct functions $\Phi_{x,y}(r, z, 0)$ and $\Phi_{x,y}(r, 0, z)$ which are regular in $|z| < 1$, continuous in $|z| \leq 1$, and satisfy (33) for every zerotuple $(\hat{z}_1, \hat{z}_2) \in A$, then $\Phi_{xy}(r, z_1, z_2)$ follows from (29) and the problem is solved. As outlined in Section 1 of [9], one may restrict oneself to consideration of a suitable *subset* $S_1 \times S_2$ of A. Indeed, if there exist curves S_1 and S_2, with $S_1 \subset \{z_1 : |z_1| \leq 1\}$ and $S_2 \subset \{z_2 : |z_2| \leq 1\}$, and a one-to-one map $z_1 = \omega(z_2)$ from S_2 to S_1 such that $(\omega(\hat{z}_2), \hat{z}_2)$ is a zerotuple of $K(r, z_1, z_2)$ for all $\hat{z}_2 \in S_2$, then the following holds. If functions $\Phi_{x,y}(r, z, 0)$ and $\Phi_{x,y}(r, 0, z)$ can be constructed that are *regular* for $|z| < 1$, *continuous* for $|z| \leq 1$, and that satisfy (33) for all zerotuples $(\hat{z}_1, \hat{z}_2)$ with $\hat{z}_1 = \omega(\hat{z}_2)$, $\hat{z}_2 \in S_2$, then *by analytic continuation* these $\Phi_{x,y}(r, z, 0)$ and $\Phi_{x,y}(r, 0, z)$ satisfy (33) for *all* zerotuples $(\hat{z}_1, \hat{z}_2)$ of A [9].

A.2 A suitable set of zerotuples

Let us now consider the construction of the curves S_1 and S_2. While there are many possible choices here, it is important to make a choice that leads to tractable numerical techniques for obtaining the analytic continuations mentioned above.

In [10], the following choice is proposed. Let s be traversing the unit circle $|s| = 1$. Let $z_1 = g(r, s)s$ and $z_2 = g(r, s)s^{-1}$, with g such that it makes the kernel zero:

$$K(r, g(r, s)s, g(r, s)s^{-1}) = 0.$$

So in our case, with the kernel given by (32), g should satisfy

$$g^2 = r\psi(gs, gs^{-1}). \tag{34}$$

Rouché's theorem implies that there are exactly two zeros of this equation satisfying $|g| \leq 1$. In fact, in our case one of these two zeros is zero, due to $\psi(0, 0) = 0$. Now take

$$S_1(r) := \{z_1 : z_1 = g(r, s)s,\ |s| = 1\},$$

$$S_2(r) := \{z_2 : z_2 = g(r, s)s^{-1},\ |s| = 1\}.$$

A.3 Preparing the ground for a Riemann boundary value problem

We next want to construct $\Phi_{x,y}(r,z,0)$ and $\Phi_{x,y}(r,0,z)$ that are regular in $|z|<1$, continuous in $|z|\leq 1$, and that satisfy (33) for all $(\hat{z}_1,\hat{z}_2)$ with $\hat{z}_1=\omega(\hat{z}_2)$, $\hat{z}_2\in S_2$. To accomplish this, one may solve a *Riemann boundary value problem*; see, e.g., Gakhov [14] for an extensive discussion of such boundary value problems, that aim to determine functions which are regular inside respectively outside a certain contour and satisfy a particular relation on that contour – the boundary. To be able to formulate such a boundary value problem, we need one more step. One can show (cf. [10], Part II) that for the curves $S_1(r)$ and $S_2(r)$ there exists a unique simple closed contour $L(r)$ in the p-plane, and functions $z_1(r,p)$ and $z_2(r,p)$ such that the following holds:

(i) $z_1(r,p): L^+(r)\to S_1^+(r)$ is regular and univalent for $p\in L^+(r)$,

(ii) $z_2(r,p): L^-(r)\to S_2^+(r)$ is regular and univalent for $p\in L^-(r)$,

(iii) $z_1(r,p)=\omega(z_2(r,p))$ for $p\in L(r)$, $\omega(\cdot)$ being a one-to-one map from $S_2(r)$ onto $S_1(r)$.

Here C^+ and C^- denote the interior and exterior of a closed contour C, and univalent means that $z_1(r,p_1)\neq z_1(r,p_2)$ for $p_1\neq p_2$. Hence (i) and (ii) can be reformulated as:

- $z_1(r,p)$ is a conformal mapping of $L^+(r)$ into $S_1^+(r)$,
- $z_2(r,p)$ is a conformal mapping of $L^-(r)$ into $S_2^+(r)$.

The curve $L(r)$ can be determined by solving a particular integral equation, cf. Section II.3.6 of [10].

A.4 The Riemann boundary value problem

Since $z_1(r,p)$ is regular and univalent for $p\in L^+(r)$ and $z_2(r,p)$ is regular and univalent for $p\in L^-(r)$, $\Phi_{x,y}(r,z_1(r,p),0)$ also is a regular function for $p\in L^+(r)$ and continuous for $p\in L^+(r)\bigcup L(r)$, and similarly $\Phi_{x,y}(r,0,z_2(r,p))$ is a regular function for $p\in L^-(r)$ and continuous for $p\in L^-(r)\bigcup L(r)$. We are now ready to formulate a standard Riemann-type boundary value problem:

Determine two functions $\Omega_1(r,p):=\Phi_{x,y}(r,z_1(r,p),0)$, $p\in L^+(r)\bigcup L(r)$, and
$\Omega_2(r,p):=\Phi_{x,y}(r,0,z_2(r,p))$, $p\in L^-(r)\bigcup L(r)$, such that $\Omega_1(r,p)$ is regular for $p\in L^+(r)$ and continuous for $p\in L^+(r)\bigcup L(r)$, $\Omega_2(r,p)$ is regular for $p\in L^-(r)$ and continuous for $p\in L^-(r)\bigcup L(r)$, $\Omega_1(r,0)=\Phi_{x,y}(r,0,0)$, $\lim_{|p|\to\infty}\Omega_2(r,p)=\Phi_{x,y}(r,0,0)$, and on the *boundary* $L(r)$, the functions $\Omega_1(r,p)$ and $\Omega_2(r,p)$ satisfy the relation (with appropriate functions $H(r,p)$ and $h(r,p)$, cf. (31)):

$$\Omega_1(r,p)=H(r,p)\Omega_2(r,p)+h(r,p),\quad p\in L(r). \tag{35}$$

The solution of this boundary value problem may be found in [14], see also Section I.2 of [10].

As observed before, now that $\Phi_{x,y}(r,z_1,0)$ and $\Phi_{x,y}(r,0,z_2)$ are found for $z_1\in S_1^+(r)$ and $z_2\in S_2^+(r)$, one obtains $\Phi_{x,y}(r,z_1,z_2)$ first for $z_1\in S_1^+(r)$ and $z_2\in S_2^+(r)$, and finally for $|z_1|\leq 1$, $|z_2|\leq 1$ via analytic continuation.

We end this appendix with several remarks.

Remark 2. It should be noticed that the initial conditions $u_{1,0}=x$, $u_{2,0}=y$ occur in the function $h(r,p)$ in (35). Further observe that, if the initial state is a set of *random variables* $(u_{1,0},u_{2,0})=(X,Y)$, with bivariate pgf $U_0(z_1,z_2)=\sum_{x=0}^{\infty}\sum_{y=0}^{\infty}P(X=x,Y=y)z_1^x z_2^y$, then the last term in the right-hand side of (29) is replaced by $z_1z_2U_0(z_1,z_2)$. This affects $h(r,p)$, but not the solution approach.

Remark 3. The kernel $K(r,z_1,z_2)$ has exactly the same form as the kernel that is considered throughout Part II of [10], but the behaviour of the random walk in the interior of the first quadrant, that we consider, slightly deviates in the interior from the behaviour of the random walk that is being considered in that Part II. That implies that a slightly different Riemann boundary value problem results.

Remark 4. It is easily seen that $\psi(0,0) = 0$; indeed, at most one customer is served per time slot, and hence random walk transitions to the South-West are not possible. In Sections II.3.10-12 of [10], kernels with this special feature are treated in detail. It turns out that now $S_1(r)$ and $S_2(r)$ are traversed *twice* if s traverses the unit circle $|s| = 1$ once. Furthermore, one has to distinguish between the cases $\frac{d}{dz}\psi(z,0)_{z=0} > (<,=)$ $\frac{d}{dz}\psi(0,z)_{z=0}$, as these cases lead to different positions of the point $z = 0$ w.r.t. the contours $S_1(r)$ and $S_2(r)$, and consequently to slightly different analytic treatments. For the sake of exposition, let us restrict ourselves to the symmetric case that the random variables $a_{1,k}$ and $a_{2,k}$ are *exchangeable*, i.e., $P(a_{1,k} = i, a_{2,k} = j) = P(a_{1,k} = j, a_{2,k} = i)$ for all $i, j = 0, 1, \ldots$, and that $\mu_1 = \mu_2 = 1$ and $\beta = \frac{1}{2}$. Then $\psi(z_1, z_2) = A(z_1, z_2)(z_1 + z_2)/2$ and $\frac{d}{dz}\psi(z,0)_{z=0} = \frac{d}{dz}\psi(0,z)_{z=0}$, leading to $z_1 = 0 \in S_1(r)$ and $z_2 = 0 \in S_2(r)$. The equation determining $g(r,s)$ now is (cf. (34)):

$$g - rA(gs, gs^{-1})\frac{s + s^{-1}}{2} = 0.$$

The contour $L(r)$ in this symmetric case turns out to be a circle, with center at $\frac{1}{2}$ and radius $\frac{1}{2}$ (see Section II.3.12 of [10]). If, moreover, only $P(a_{1,k} = a_{2,k} = 0) = A_{0,0}$, $P(a_{1,k} = 1, a_{2,k} = 0) = A_{1,0}$, $P(a_{1,k} = 0, a_{2,k} = 1) = A_{0,1}$ and $P(a_{1,k} = 1, a_{2,k} = 1) = A_{1,1}$ are possibly non-zero, then the equation determining $g(r,s)$ reduces to a quadratic equation:

$$g - r[A_{0,0} + A_{1,0}g(s + s^{-1}) + A_{1,1}g^2]\frac{s + s^{-1}}{2} = 0,$$

from which the contours $S_1(r)$ and $S_2(r)$ are easily determined.

Remark 5. If the steady-state distribution of the joint queue-length distribution in our model exists, then its generating function $\Phi(z_1, z_2)$ can be determined via $\Phi(z_1, z_2) = \lim_{r\to 1}(1 - r)\Phi_{x,y}(r, z_1, z_2)$. In Section II.3.9 of [10], the details of this approach are presented; in Section II.2.16, for the case of a *symmetric* two-dimensional random walk, it is outlined how one can *directly* handle the steady-state case. Here one considers the kernel $K(1, z_1, z_2)$, identifying a contour $L(1)$ and zeros $(z_1(1,p), z_2(1,p))$ of the kernel, and formulating a Riemann boundary value problem for $\Phi(z_1(1,p), 0)$ and $\Phi(0, z_2(1,p))$ with boundary $L(1)$.

Remark 6. In order to determine performance measures like the mean steady-state queue lengths, one has to evaluate quantities like $\frac{d}{dz}\Phi(z,1)_{z=1}$. Using our approach, this requires the numerical determination of $L(1)$, and of (the analytic continuation of) the conformal mappings from $S_1(r)$ and $S_2(r)$, respectively, to $L(r)$. Finally, a numerical analysis is required of the singular contour integrals that specify $\Phi(z,0)$ and $\Phi(0,z)$. We refer to Part IV of [10] for an extensive discussion of such numerical aspects. While it turns out to be possible to obtain numerical values of such performance measures, the analysis is quite involved. We have hence decided to concentrate in this paper on an alternative approach for the *steady-state* analysis of the queueing model under consideration – a method that has several novel aspects, and that allows us to obtain numerical values in a more straightforward manner. Finally, let us mention that these numerical problems could be circumvented by using the approach in [13] that strongly relies on analytic continuations, and that makes use of Weierstrass elliptic functions or a Fredholm integral equation instead of a conformal mapping.

B Proof of analyticity of U in a neighborhood of $\beta = 0$

In this appendix, we argue that $U(z_1, z_2; \beta)$ is an analytic function in a neighborhood of $\beta = 0$. We therefore use the implicit function theorem for Banach spaces (see theorem (10.2.3), page 272 in [11]), stated first.

Lemma 1 (An implicit function theorem for Banach spaces)**.** Let E, F, G be three Banach spaces, f a p-times continuously differentiable mapping of an open subset A of $E \times F$ into G. Let (x_0, y_0) be a point of A such that $f(x_0, y_0) = 0$ and that the partial derivative $D_2 f(x_0, y_0)$ is a linear homeomorphism of F onto G. Then there is an open neighborhood Y_0 of x_0 in E such that, for every open connected neighborhood Y of x_0, contained in Y_0, there is a unique continuous mapping u of Y into F such that $u(x_0) = y_0$, $(x, u(x)) \in A$ and $f(x, u(x)) = 0$ for any $x \in Y$. Furthermore, u is p-times continuously differentiable in Y.

The proof of analyticity of $U(z_1, z_2; \beta)$ in a neighborhood of $\beta = 0$ evolves as follows. First, define B_2 as the Banach space comprising all bivariate analytic bounded functions in D^2, with D the open complex unit disk. Similarly, define B_3 as the Banach space comprising all trivariate analytic bounded functions in D^3, that have a limit of 0 for the first two arguments going to 1.

Define E, F and G in Lemma 1 as $E = \mathbb{C}$, $F = B_2$ and $G = B_3 \times \mathbb{C}$.

We define the mapping f as:

$$\begin{aligned} f(\beta, U) =& [K(z_1, z_2, \beta)U(z_1, z_2) - K_{00}(z_1, z_2, \beta)U(0, 0) \\ & - K_{10}(z_1, z_2, \beta)U(z_1, 0) - K_{01}(z_1, z_2, \beta)U(0, z_2), U(1, 1) - 1] \end{aligned}$$

with K, K_{00}, K_{10} and K_{01} given in (3)-(6).

This mapping is defined from an open subset A of $E \times F$ to G, where A includes the point $(0, V_0)$ (V_0 is the pgf of the system content of both classes for the HOL priority scheduling, i.e. for $\beta = 0$).

Since K, K_{00}, K_{10} and K_{01} are bounded analytic functions in D^3, and since f is affine in U and β, it is easily seen that f is p-times continuously differentiable for all p. We further observe that

$$f(0, V_0) = [0, 0].$$

Since f is affine in U, the (Banach space) derivative $d_2 f(0, V_0)$ (see [11], section 8 for more explanation on differential calculus in Banach spaces) equals

$$d_2 f(0, V_0)(U) = [K(z_1, z_2, 0)U(z_1, z_2) - K_{00}(z_1, z_2, 0)U(0, 0) - K_{10}(z_1, z_2, 0)U(z_1, 0), U(1, 1)].$$

We now prove that this mapping is a homeomorphism in four successive steps, namely that (i) it is continuous, that the mapping is (ii) injective and (iii) surjective (and thus bijective) and that (iv) the inverse $(d_2 f(0, V_0))^{-1}$ is a continuous mapping.

- First of all, it is clear that $d_2 f(0, V_0)$ is a continuous mapping for the same reasons as that the mapping f itself is continuous.
- Assume that $d_2 f(0, V_0)(U_1)$ equals $d_2 f(0, V_0)(U_2)$ for given U_1 and U_2. Then

  $$\begin{aligned} K(z_1, z_2, 0)(U_1(z_1, z_2) - U_2(z_1, z_2)) - K_{00}(z_1, z_2, 0)(U_1(0, 0) - U_2(0, 0)) & \\ - K_{10}(z_1, z_2, 0)(U_1(z_1, 0) - U_2(z_1, 0)) = 0, & \\ U_1(1, 1) - U_2(1, 1) = 0. & \end{aligned}$$

  Or in other words,

  $$f(0, U_1 - U_2) = (0, -1).$$

  However, this functional equation has the zero-solution as a unique solution (see Theorem 3.3 in [5]), and as a result $U_1 = U_2$ and $d_2 f(0, V_0)$ is injective.
- To prove that the function $d_2 f(0, V_0)$ is surjective, we solve the equation $d_2 f(0, V_0)(U) = (g, c)$ with g a bivariate analytic bounded function in D^2 with limit 0 for its arguments going to 1, and c a complex number. We thus solve

  $$\begin{aligned} & K(z_1, z_2, 0)U(z_1, z_2) - K_{00}(z_1, z_2, 0)U(0, 0) - K_{10}(z_1, z_2, 0)U(z_1, 0) = g(z_1, z_2), \\ & U(1, 1) = c. \end{aligned}$$

  This functional equation has the same form as functional equation (9), and the same technique can be applied, leading to the solution

  $$U(z_1, z_2) = \left[\left(1 - \frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2} \right) c - \frac{g^{(1)}(1, 1)}{\mu_1} - \frac{g^{(2)}(1, 1)}{\mu_2} \right]$$

$$\times \frac{K_{00}(z_1,z_2,0)K_{10}(z_1,Y(z_1),0) - K_{10}(z_1,z_2,0)K_{00}(z_1,Y(z_1),0)}{K_{10}(z_1,Y(z_1),0)K(z_1,z_2,0)}$$
$$+ \frac{g(z_1,z_2)}{K(z_1,z_2,0)} - \frac{K_{10}(z_1,z_2,0)g(z_1,Y(z_1))}{K_{10}(z_1,Y(z_1),0)K(z_1,z_2,0)},$$

with $g^{(j)}(1,1) = \partial g(z_1,z_2)/\partial z_j|_{z_1=z_2=1}$ and

$$Y(z_1) = \frac{A(z_1,Y(z_1))\mu_2}{(1-A(z_1,Y(z_1))(1-\mu_2)}.$$

- The U obtained in the previous bullet is $(d_2 f(0,V_0))^{-1}$. This mapping is easily seen to be continuous.

We conclude that the mapping $U \to d_2 f(0,V_0)(U)$ is a linear homeomorphism. It thus follows from Lemma 1 that $U(z_1,z_2;\beta)$, as defined in the paper, is p-times differentiable at $\beta = 0$, and this for all p.

Remark 7. Although the particular two-dimensional process in this appendix, and hence our Riemann-Hilbert problem, is a special case of the more general class of problems treated in [10], there is an alternate approach to solving the functional equation (31) by exploiting the special structure of our problem. From (31) we see that

$$\beta K_{10}(z_1,z_2) + (1-\beta)K_{01}(z_1,z_2) = 0, \tag{36}$$

so that the two functions K_{10} and K_{01} are equal up to a multiplicative constant. This feature makes that the Riemann-Hilbert problem can be reduced to a Dirichlet problem for a circle, for which Schwarz's formula yields the solution to the functional equation in terms of an elliptic integral. Such integrals provide efficient algorithms for computing performance characteristics by numerical integration. This approach was taken in [8, 12, 15]. We choose not the pursue this approach here, as it requires a construction based on the analytic continuations of the functions K_{10} and K_{01}, and a detailed study of the involved conformal mappings, which is rather involved.

Ackowledgment The first author wishes to thank Dieter Fiems for stimulating conversations and for pointing into the direction of Banach spaces for the analiticity proof in Appendix B. We thank the referee for his/her detailed reading of the manuscript and constructive comments.

References

[1] I.J.B.F. Adan, O.J. Boxma, and J.A.C. Resing. Queueing models with multiple waiting lines. *Queueing Syst.*, 37(1-3):65–98, 2001.

[2] I.J.B.F. Adan, J.S.H. van Leeuwaarden, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[3] I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm. A compensation approach for two-dimensional Markov processes. *Advances in Applied Probability*, 25(4):783–817, 1993.

[4] J.R. Artalejo, I. Atencia, and P. Moreno. A discrete-time $Geo^{[X]}/G/1$ retrial queue with control of admission. *Applied Mathematical Modelling*, 29(11):1100–1120, 2005.

[5] S. Asmussen. *Applied Probability and Queues*. Wiley, New York, 1987.

[6] J.P.C. Blanc. On a numerical method for calculating state probabilities for queueing systems with more than one waiting line. *Journal of Computational and Applied Mathematics*, 20:119–125, 1987.

[7] J.P.C. Blanc. A numerical study of a coupled processor model. In *G. Iazeolla, P.-J. Courtois, O.J. Boxma (eds.), Computer Performance and Reliability*, pages 289–303, 1988.

[8] A. Brandt and M. Brandt. On the sojourn times for many-queue head-of-the-line processor-sharing systems with permanent customers. *Math. Methods Oper. Res.*, 47(2):181–220, 1998.

[9] J.W. Cohen. Boundary value problems in queueing theory. *Queueing Systems Theory Appl.*, 3(2):97–128, 1988.

[10] J.W. Cohen and O.J. Boxma. *Boundary Value Problems in Queueing System Analysis*. North-Holland, Amsterdam, 1983.

[11] J. Dieudonné. *Foundations of Modern Analysis*. Academic Press, New York, 1969.

[12] G. Fayolle and R. Iasnogorodski. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(3):325–351, 1979.

[13] G. Fayolle, R. Iasnogorodski, and Malyshev V. *Random Walks in the Quarter Plane: Algebraic Methods, Boundary Value Problems and Applications*. Springer, Berlin, 1999.

[14] F.D. Gakhov. *Boundary Value Problems*. Pergamon Press, Oxford, 1966.

[15] F. Guillemin and D. Pinchon. Analysis of generalized processor-sharing systems with two classes of customers and exponential services. *J. Appl. Probab.*, 41(3):832–858, 2004.

[16] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal of Applied Mathematics*, 48(5):1159–1166, 1988.

[17] J.F.C. Kingman. Two similar queues in parallel. *The Annals of Mathematical Statistics*, 31(4):1314–1323, 1961.

[18] A.G. Konheim, I. Meilijson, and A. Melkman. Processor-sharing of two parallel lines. *Journal of Applied Probability*, 18(4):952–956, 1981.

[19] S.G. Krantz and H.R. Parks. *The Implicit Function Theorem: History, Theory and Applications*. Birkhäuser, Boston, 2002.

[20] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf. Analysis of cycle stealing with switching times and thresholds. *Performance Evaluation*, 61(4):347–369, 2004.

[21] J.A.C. Resing, G. Hooghiemstra, and M.S. Keane. The M/G/1 processor sharing queue as the almost sure limit of feedback queues. *Journal of Applied Probability*, 27(4):913–918, 1990.

[22] W.B. van den Hout. *A Numerical Approach to Markov Processes*. PhD Thesis, Tilburg University, Tilburg, The Netherlands, 1996.

[23] J. Walraevens, B. Steyaert, and H. Bruneel. Delay characteristics in discrete-time GI-G-1 queues with non-preemptive priority queueing discipline. *Performance Evaluation*, 50(1):53–75, 2002.

[24] J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline. *European Journal of Operational Research*, 157(1):130–151, 2004.